DOCUMENT RESUME

ED 293 878 TM 011 468

AUTHOR Nasca, Donald
TITLE An Educators' Field Guide to CRT Development and Use in Objectives Based Programs.
PUB DATE 17 Mar 88
NOTE 42p.
PUB TYPE Information Analyses (070) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Academic Achievement; *Criterion Referenced Tests; Elementary Secondary Education; Mastery Tests; Measurement Techniques; Norm Referenced Tests; Objectives; *Test Construction; Test Use

ABSTRACT

A basic understanding of criterion-referenced test (CRT) use and development is presented with definitions and characteristics of CRTs. The steps necessary to construct and validate a CRT, the appropriate use of CRTs, the historical development of CRTs, and terms used in conjunction with CRTs are discussed. The most common definition of a CRT is a sample of items yielding information directly interpretable with respect to a well-defined domain of tasks and specified performance standards. The major differences from norm-referenced tests are in development and interpretation. If a test has not been normed and the test items assess performance specified in objectives, it is a CRT. CRTs test only what has been taught and are curriculum aligned. CRTs are used to obtain examinee scores with some absolute meaning relative to a district curriculum. Validity, reliability, discrimination, and the degree of difficulty of CRTs are discussed. CRTs must be field tested on repeated occasions if initial use produces undesirable statistical properties. The possibility of repeated testing should be built into any plan for CRT development. A 38-item reference list is included. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED293878

An Educators' Field Guide to CRT Development and Use in Objectives Based Programs.

March 17, 1988

Donald Nasca Ed.D

Professor of Educational Administration
State University College at Brockport
Brockport NY, 14420

Abstract

A confusing array of terms used to describe Criterion-Referenced tests and their development has led to this attempt at reconciling meaning and interpretation. This paper describes the steps necessary to construct and validate a CRT, describes appropriate use of CRT's, reviews the historical development of CRT's and clarifies the terms used in conjunction with CRT's. A substantial bibliography citing both technical and popular sources is included.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

DONALD NASCA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 011 468

BEST COPY AVAILABLE

An Educators Field Guide to CRT Development and Use in Objectives-Based Programs

Although the term Criterion-Referenced Measure was coined by William Glaser in a paper in 1962 (Glaser & Klaus, 1962) and substantial rhetoric has appeared in the literature since that date, no single source has attempted to bring together the practical applications and technical features of CRT's for examination by educational practioners. Lathrop (1986), for example, concluded that, "Although many technical articles have been written concerning the reliability of competency tests, few practioners appear to find such discussions helpful." p. 234. Discussions of validity and CRT item construction have been similarly unenlightening.

A variety of terms with essentially similar meanings have emerged in conjunction with Criterion-Referenced Testing. Programs with such varied titles as Competency Based, Learning for Mastery, Data Based Instruction, Mastery Learning, Group Based Mastery Learning and Outcome Based are all grounded in CRT technology. A lack of uniformity and array of terms with only minor variations in definition has created a somewhat confusing picture of CRT application.

This paper is an attempt to clarify CRT terminology as well as offering to practioners a basic understanding of the current state of CRT use and development. It begins with basic definitions and characteristics of CRT's, describes development of CRT's, discusses the technical issues related to CRT's in laypersons' terms , summarizes successful applications of CRT use and offers a suggestion for solving one of the more controversial issues surrounding CRT use. A bibliography with both popular and technical sources is referenced.

Basic Definitions

The term, Criterion-Referenced Measurement was originally defined as a measure of student performance on a hierarchially arranged achievement continuum. This continuum was organized around psychological and developmental variables associated with each content area and maximized the probability of identifying a student's skill level within the pre-defined continuum. The student's test score was then interpretable as evidence of skill attainment within the hierarchy and provided direct evidence of skills mastered with implications for perscriptions for future development. This perspective on achievement testing was introduced in response to dissatisfaction with the almost exclusive reliance on 'Norm-Referenced' measures that identified a student's skill level in terms of 'Normal'

development of students with similar educational experience. Norm-Referenced measures ignore the concept of an 'Achievement Continuum' and focus rather on 'Population' performance. The score obtained on a Norm-Referenced measure may be interpreted only as a comparison with the performance of other, similar individuals. This score has little personalized diagnostic or perscriptive value.

Criterion-Referenced Tests have lost some of their original intent in current usage and rather than focusing on an 'Achievement Continuum', tend to focus on specific programs, curricular sequences and/or objectives. This modified adaptation of CRT's has given rise to the terms Curriculum Referenced Test and/or Program Referenced Test. Technically speaking, Curriculum Referenced is a more appropriate designation for many of the current school testing applications. Because CRT's have been modified in current applications, they no longer contain the precise characteristics originally proposed by Glaser and although there remains considerable similarity between CRT's and Curriculum Referenced tests, if one wishes to be technically precise, one would distinguish between the two terms. The major difference between CRT's and Curriculum Referenced is the source of objectives. CRT's begin with a psychological and/or developmental continuum of skills while Curriculum Referenced Tests begin with specific courses of study. One

could easily argue that courses of study are psychologically and/or developmentally arranged hierarchies and that the distinction between CRT and Curriculum Referenced is relatively meaningless. This paper continues to use the term CRT while describing curriculum referenced applications.

The most currently common definition of CRT is, "... a sample of items yeilding information that is interpretable directly with respect both to a well defined domain of tasks and to specified performance standards." (Tindall, et. al. 1985, p. 203) The tasks for which performance standards are specified are derived from Competencies or Behavioral Objectives. These are the educational goals obtained from curriculum guides and/or written by district personnel. There is little, except semantic preference, to distinguish between competency and objective. Both define educational goals in terms of learner performance and both require elaboration for translation into CRT items.

The tasks most frequently subjected to CRT scrutiny are called Endpoint, Terminal or Outcome objectives. That is, the major outcomes in a grade/subject are defined and converted into student performance statements. Again, there appears to be no distinguishing differences among the terms, 'Endpoint', 'Terminal' and 'Outcome'.

Mastery, Competence and Proficiency are words used interchangable to describe student attainment of behaviors implied in the objectives statement. Again, use of these terms is subject only to semantic preference.

Endpoint or terminal objectives are often further defined by Enabling or Perequisite objectives. Enabling objectives are derived from a Task Analysis of the terminal objective. That is, each endpoint objective is broken down into its component behaviors. These enabling objectives are often considered the major focus of instruction with decisions on how best to present and attain these perequisite behaviors left up to teachers while endpoint objectives are the end of instruction standards set by a school district. Administrators advocating the adoption of CRT technology frequently state, "We are not telling you how to teach, only what to teach." or "We have specified the endpoint behaviors, its up to teachers to determine how best to get there.".

CRT Characteristics

It is impossible to distinguish between norm-referenced and CRT items based merely on appearence. Exactly the same items could appear on either test. The major differences are in development and interpretation. If the test has been "Normed" then interpretation of scores is based on

comparisons with other individuals of similar age and educational experience. If the test has not been normed AND the test items assess performace specified in objectives, then it is a CRT.

CRT's reflect only the educational program adopted by the district. Norm-referenced tests, on the other hand, assess the average of all state curriculum guides and textbook publishers scope and sequence charts. McGraw-Hill (ORBIT, 1984) has stated in the rationale for its support of CRT's that, "... a growing trend toward individualized, objectives-based instruction has uncovered a need for a measure of student performance relative to curriculum that is more precise than that afforded by norm-referenced tests." p. 1.

CRT's posses Curricular Alignment. They test only what has been taught based on an assumption that whatever gets measured, gets taught. This raises a question of the stated objectives becoming the minimum program and perhaps restricting the development of higher level and creative thought processes. This could, of course happen unless higher level and creative processes are specified as outcome behaviors.

Because CRT's are aligned with the curriculum, they have the capacity to provide diagnostic information for teacher use

in instructional grouping decisions. One body of research has demonstrated that frequent monitoring of students in highly focused instructional settings produces significant achievement gains. Although these findings have more recently been challenged there remains a considerable body of evidence supporting the use of frequent student monitoring.

CRT Use

The question often arises, "Should I be using a CRT or a norm referenced test?". The answer is simply, "What purpose do you wish to serve?" If the purpose is to compare examinee performance with other, similar individuals; then norm referenced. If the purpose, on the other hand, is to obtain an examinee score that has some absolute meaning relative to a district curriculum; then CRT. Districts may wish to use both a norm-referenced test to compare student performance with national norms and CRT's to identify specific student needs in terms of district curricula. If over testing becomes an issue, norm-referenced tests may be used with only a sample of the student population annually and still provide an accurate estimate of district achievement parameters. That is, alternate grades or classrooms may be tested with norm-referenced tests and still provide accurate district norms.

This does bring up the question of testing in general. Testing is carried out either to demonstrate accountablity and/or to provide information for decision making at the individual classroom, building and/or district level. Practically, testing is designed to produce information for decision making and as long as test data is legimately being used to make decisions, then that testing is necessary and appropriate. School boards often insist on norm-referenced tests to demonstrate accountability. Seldom is this data used for decision making and may, in fact, interfere with more appropriate diagnostic data collection.

CRT's may be used for diagnostic purposes during the course of daily instruction, they may be used for end of year placement/promotion decisions, as a graduation requirement and/or licensure decisions. Although much of what follows focuses more specifically on CRT use in diagnostic and placement decisions, generalizations are easily made to other uses.

CRT's are extremely flexible to use. Items assessing one objective may be administered whenever the teacher desires and/or several competencies may be assessed in one setting. CRT data is useful both in Formative evaluation (information obtained on students for the purpose of making instructional, diagnostic decisions) and/or Summative evaluation (end of year data used in making placement

decisions). This flexibility gives CRT's two primary functions. First, short tests (4 to 5 items) may be used for estimates of specific objective mastery that may then be used in classroom instructional management decisions and second, an end of year test based on a collection of items may be used for placement decisions.

District developed CRT's may be printed on machine scorable pages and scanned and analyzed either within the building on microcomputers or scanned and analyzed centrally. Technological developments in Instructional Management Systems (Sheppard, 1986; Witthuhn, 1986)) has contributed significantly to the use of formal testing in instructional decision making by providing immediate turn around of test results. This immediate turn around feature used to be available only with teacher made tests that were time consuming to construct and correct and lacked both reliability and validity data.

The relationship between test items appearing on formative measures and items appearing on summative measures introduces an unresolved dilemma. If teachers are directed to focus instruction on specific objectives for which formative measures have been developed and are being used to monitor instruction, then should those same items appear on the summative measure? Or, should the summative measure represent a more global assessment of instruction while

formative measures assess developmental and/or enabling objectives? If one could be sure that each of the parts (formative) added up to the whole (summative) then it would be safe and reasonable to use formative measures for developmental purposes and summative measures for more global purposes. For example, reading skill development would become the focus of formative evaluation while a more wholistic measure of reading comprehension would serve as the summative measure. This relationship still needs to be evaluated before definitive decisions can be reached and it is entirely possible that the final decision will vary across disciplines and levels.

CRT's may also be used as pre-tests to provide a survey of student skills at the beginning of an instructional sequence. CIMS math, for example, makes extensive use of survey tests (CIMS, 1986) at the beginning of the year to assist in instructional planning.

Development

The development of CRT's begins with objectives that have been written or selected by district curriculum specialists. Objectives are more easily communicated if they are simple statements of student outcomes. For example; "Students will convert word problems into number sentences." or "Students will draw inferences from grade level reading selections."

This approach is preferred to that of earlier writers (eg. Mager, Popham, etc.) who recommended that each objective contain not only the behavior to be achieved but also the condition and criterion for demonstrating attainment of that behavior. This led not only to extremely awkward and lengthy statements but required several different skills that are more appropriately apportioned to different persons.

Experience has demonstrated that something in the neighborhood of 20 well defined objectives at a relatively high level of generality is a reasonable number per subject per grade level. This translates into 80 test items on a summative test if one uses the criteria of 4 test items per objective. Each of these higher level 'terminal' objectives may be broken down into perequisite objectives for instructional and formative evaluation purposes.

Once agreement on objectives has been reached, overtly observable student responses must be defined for each objective. This observable student response is labeled Condition and identifies the parameters of a knowledge or content Domain from which test items may be drawn to assess mastery status of each objective. There are two approachs to defining conditions; Deductive and Inductive.

1. A Deductive approach begins with a narrrative statement that clearly defines the parameters of a

domain. The statement must make it possible to discriminate between appropriate and inappropriate test items.

Example:

Objective - (Third grade arithmetic problem solving): Students will convert word problems into number sentences.

Condition: Given a three sentence, one step word problem requiring either addition or subtraction with no extraneous information and single digit values, students will select a number sentence representing the information presented in the problem. (Note - the criteria for determining mastery need not be stated in the condition. That is, the number of correct responses required for demonstrating mastery need not be included at this point.)

2. The Inductive approach begins with examining a variety of test items. These items are divided into two groups; those that reflect the intent of the objective and those that do not reflect the intent. From these two groups, it is then possible to write a condition statement.

Express Intent		Do Not Express Intent	
2 3	9 6	2 3 x 4 6 =	9 6 x 5 =
x 4 6	x 5		
		Any item with multiple choice answers.	
		8 6 X 1 0 5 =	

Condition: Given multiplication examples in a vertical format with two digit multiplicands and one or two digit multipliers, students will compute and write the product.

An alternative condition statement at a more general level is:

Given two digit multiplication examples in either a vertical or horizontal format, students will compute and either write or select the correct product.

The level of specificity at which a condition is written may be a function of whether the test items are to be used for

formative or summative purposes. Items at the formative level might well be more specific.

Although the step of condition writing must be directed by a measurement specialist, it is desirable to involve classroom teachers in the process. The range of expectations and perspectives offered by classroom teachers with respect to the specific population they are dealing with adds substantially to final acceptance of the CRT approach. It is also desirable to stress that objectives have not been written in stone and that in order to develop conditions for overtly observable responses, it may be necessary to modify some objectives.

Measurement experts agree that specification of the domain from which test items may legitimately be drawn for evaluation of objectives is considered the single most important step in the construction of CRT items. Hambleton & Novick (1973) for example, state, "If the proper domain of test items measuring an objective is not clear, it is impossible to select a representative sample of test items from that domain." p. 32

The importance of this step is illustrated in the following example:

Objective: "Students will be able to describe action portrayed in a picture."

Focus of instruction by:

Teacher #1: Teaches students to isolate details in the picture and to relate each of these details in a sentence.

Teacher #2: Teaches students to infer what happened immediatly prior to and after the picture was taken and to describe the hypothetical sequence of events in a three sentence paragraph.

Teacher #3: Teaches students to relate the picture to personal experience and to desribe how people in the picture might feel about what is happening.

Condition: Given a picture portraying action, students will be able to write a three sentence paragraph accurately describing the action.

Although each intrepretation of the objective as given is accurate, each is different and the "test item" fails to assess the focus of instruction by any of the three teachers. Because of these very typical differences in interpretation of an objective, it is imperative that a Knowledge domain for each objective be clearly defined.

Although a substantial proportion of conditions will specify traditional paper and pencil type responses, some objectives

may more appropriately call for student generated oral and/or written responses. For example:

Read fluently a 150 word passage in 2 minutes.

Pronounce consonant sounds.

Answer a question in a complete sentence.

These are called Teacher Certified competencies and each requires clearly defined criteria to discriminate between correct and incorrent responses. Experience has demonstrated that there will be considerable teacher variability in rating student generated responses and teacher certified competencies have a potential for extremely low reliability. Experience has also shown that these items present excellent opportunities for in-service training in which student performance standards in response to these open ended items are the foci of discussion.

Once conditions defining the parameters of knowledge/content domains for each objective are specified, test items from each domain are generated and a random set of these items selected for use in assessing student mastery. Test item construction is a technical task requiring the assistance of a measurement specialist. Both textbook unit tests and standardized tests may be used as models for item construction.

Although the number of items needed for an accurate estimate of total domain performance can be determined statistically, these techniques are beyond the scope of this paper and readers are refered to Haladyna & Roid (1983) and Hambleton, Craig & Simon (1983) for detailed steps. In actual practice, it is common to find 4 to 5 items for each objective. This number will vary depending on how broadly or narrowly the objective is written. Generally, it is desirable to focus each objective on such a narrow range of behaviors that variabilty within the knowledge domain of a given objective is minimal. This does, however, raise another dilemma. Too narrowly focused objectives raise the issue of Specific Sterility and may limit transfer while too broadly defined objectives lead to ambiquity in the instructional focus. Developers are warned of this problem and encouraged to work toward compromises.

If CRT's are to be used both formatively and summatively and/or alternate forms of tests are to be used, it will be necessary to develop a pool of several test items for each objective. The relationship of formative to summative measures as indicated earlier, will also influence the task of test item construction.

Several test publishers and educational organizations have created banks of CRT items that may be purchased once domain specifications have been developed. Publishers are often

anxious to identify appropriate test items for districts that have specified their student response conditions.

Once test items from each domain have been selected or created, these test items must be field tested. This leads to several technical considerations discussed in the next section.

Validity, Reliability, Discrimination and Degree of Difficulty

Before discussing in detail each of these attributes of CRT item construction, it should be understood that standard techniques used in norm referenced test construction must be modified for CRT use. The primary reason for this is the lack of variability in student responses that occurs in CRT's. That is, the basic statistical procedures used in norm referenced test constructioon are dependent upon substantial variability in student responses to test items. In fact, norm referenced test items are constructed so as to maximize this variability. Popham & Husek (1969) state that, "With criterion-referenced tests, variability is irrelevant." p. 3 and Hambleton et. al. (1978) concluded that, "... classical approaches to reliability and validity estimation will need to be interpreted more cautiously (or discarded) in the analysis of criterion-referenced tests." p. 15

A second factor impacting on statistical data supporting a CRT is the seriousness of decisions reached as a result of interpreting scores. The statistical support for a CRT used as a high school graduation requirement needs to be far more rigorous than for a CRT used during the course of instruction for grouping decisions. In the later case,

additional information is generally available to the teacher and there is ample opportunity to review each decision.

Validity Validity is technically not a characteristic of a test but rather a consideration related to the inferences drawn from an examinee's test score. That is, a test of two digit multiplication is valid for inferences relative to two digit multiplication. It is not valid for inferences related to either one digit or three digit multiplication.

Validity has often been described as though there were several different forms of validity, eg. content, construct, concurrent, face, logical or predictive. In reality, there is only one form of validity and that involves the relationship of a decision based on test scores to a true state of being. That is, an inference based on test performance is being made about the true or real nature of an examinee. At best, test performance is only an estimate of an individuals capability and validity is an attempt to quantify the accuracy of this estimate. There are multiple approaches to this quantification process; i.e. content, construct, etc.

CRT scores are generally used for two different kinds of inferences (decisions) and therefore require two different types of validity.

The first validity issue is: 'Does the score obtained by a student on a sample of test items accurately reflect total domain performance?' Stated differently, 'Does student performance on the sample of test items provide an accurate estimate of the total domain score?' This is called Domain Score Validity. The domain of interest may be a single objective as illustrated in an earlier example or it may be something as extensive as reading comprehension. Domain score validity is more of a concern when examining test items related to a specific objective, that is, formative evaluation.

The second validity issue is: 'Has the student mastered the domain(s) from which test items have been drawn?' Again, stated differently, 'Does the test accurately discriminate between students who have mastered the objective(s) and those who have not mastered the objective(s)?' This is called, Mastery Status Validity. Mastery status validity is more of a concern with summative tests that contain a range of domains covered during the year.

Domain Score Validity may be established by comparing student performance on a sample of test items drawn from a domain with performance on all test items included within that domain. For example, in the objective: "Name the numerals 0 - 9", a sample from the domain might include only four test items while the entire domain contains ten

possible responses. A statistical comparison (correlation) of student performance on the sample of four items with performance on the total domain of ten items would provide a statistical value for this domain score validity. This is an easy task when the domain is well defined, discrete and relatively small. Reasonable limitations are quickly exceeded when one considers that the domain for pairs of two digit multiplication examples in a vertical format contains 8,100 possible different responses and that the domain for the objective; 'Drawing inferences from a reading selection.' is indeterminate. It would be impossible to assess student performance on the total domain in either of these latter examples.

In actual practice, domain score validity is most often established by a process involving the use of "Judges" who examine test items to determine if they accurately reflect the domain. Agreement across several judges is used to establish 'Face' or 'logical' validity of the domain inference. It is assumed that if several judges agree that a sample of test items accurately represents the domain, then performance on the sample is an accurate estimate of the domain score.

Mastery Status Validity is a far more difficult issue and by far the most controversial adaptation of CRT's. Technically, mastery of a domain implies that a student will be able to

respond correctly to every test item from that domain. Realistically, because of guessing, lapses in attention, carelessness and measurement error, standards for determining domain mastery are generally set at something less than 100%. It is common to find cut-off scores in the 70 to 80% range. Several different approaches have been used to establish a 'standard' (cut-off score) that accurately discriminates between masters and non-masters. Examples include:

1. **Use of Judges** - Expert judges examine the competencies and test items and predict the performance of a borderline student on each test item. The probabilities of borderline student responses on individual items are summed to determine a score for the boundary between Mastery and Non-Mastery. Secolsky (1987) has, however, demonstrated that there is considerable variability in expert judgment.

   The Cut-Off score (standard) arrived at in this manner means that the student has mastered what a group of judges believes to be the performance of a minimally qualified student. The validity statistic is the correlation across all judges.

2. **Teacher Prediction of Mastery** - Teachers familiar with each student predict who the Masters are and

these predictions are compared with test performance. A Cut-Off score is then established to include as large a proportion of teacher predicted masters above the cut-off as possible and as large a proportion of teacher predicted non-masters below the cut-off as possible.

Masters determined by this approach are students whose minimum score is like that of successful students in past years based solely on teacher judgment. The validity statistic is the correlation between teacher judgment and student test performance.

3. **Predictive Capacity** - CRT performance is correlated with some future measure of performance and this correlation is used to determine the CRT score necessary to insure future success.

Masters determined by this approach are students who are similar to those who have experienced continued success in the past. The validity statistic is the correlation between current and future performance.

Each of these standard setting procedures will produce a different cut-off score and there is no substantive defense for use of any of these procedures. The use of an absolute cut-off score, that is, a pre-determined score for use in determining mastery status, can only create controversy. The

development and use of a relative cut-off score will be discussed later.

Regardless of where a standard is set, there will be errors created because test items do not adequately reflect the domains assessed (validity) and because student performance on the test varies from one occasion to the next (reliability). The greatest number of errors will occur closest to the cut-off score regardless of which standard setting procedure is used. That is, students scoring just below or just above the cut-off score are most likely to be missclassified. These are true masters who score below the cut-off and true non-masters who score above the cut-off. (Fig. 1)

Fig 1.

Cut-off Score Errors

Test Estimate	True Ability	
	Masters	Non-Masters
Masters		Error
Cut-off score		
Non-Masters	Error	

Although a number of rather complex statistical procedures have been put forth in an attempt to reduce these errors, (Hambleton, 1978; Hambleton & De Gruijter, 1983; Haladyna &

Roid, 1983; Linn, 1978 & Shepard, 1980;) they are generally overly cumbersome for the results obtained and each can do little more than reduce the number of errors. The errors cannot be eliminated!

The standard setting limitation of CRT use has received substantial criticism and has led at least one leading measurement expert (Glass, 1978) to conclude that CRT's should not be used for mastery/non-mastery classifications but instead, should be used only to determine if the rate of learning goes up or down. Educational placement decisions are then attached to a rate of learning interpretation. A host of other experts, on the other hand, (Berk, 1980; Hambleton, 1978; Popham & Husek, 1969 and Shepard, 1980) offer that the arbitrary standard imposed by CRT's is better than no standard and certainly the achievement gains attributed to CRT use (Abrams, 1985; Guskey, et. al. 1986 and Fuchs, et. al. 1986) suggest that there is substantial value in the use of CRT driven instruction.

An alternative to use of an absolute cut-off score for determining mastery status that has significant potential for reducing missclassification errors has been introduced by Lathrop (1986). His approach calls for two cut-off scores with an 'uncertain' area between these two scores. (Fig. 2) Students above the upper cut-off score are clearly masters while students below the lower score and clearly

non-masters. Decisions on students in the 'uncertain' area can be based on a variety of data. Lathrop recommends additional testing. If, however, formative evaluation during the year has been recorded, this would appear to have great value in determing mastery/non-mastery status of students in the 'uncertain' area.

Fig. 2

Cut-off Score Range

	True Ability	
Test Estimate	Masters	Non-Masters
Masters		Error
Top Cut-off score		
Uncertain	Historical information added to make final decisions	
Bottom Cut-off score		
Non-Masters	Error	

This approach has significant potential for use in situations where CRT's are being used for grade level promotion. Particularly in view of recent evidence that straight non-promotion appears to be counterproductive. (Peterson, et.al. 1987; Holmes & Matthews, 1984) That is, students failing to meet minimum competency requirements

seldem benefit from repeating an already proven failure experience. Although there are students in need of remediation, that remediation is more profitably presented in the form of an alternative and the specific alternative required could more easily be determined given CRT performance data.

Relative Cut-Off Score: If one accepts the assumption that educational outcomes are at least partially a function of available resources and that educational resource allocations are not made based on educational need but rather on political and economic consideration, then it follows that resource allocation should serve as the basis for mastery/non-mastery decisions. The effectivness of any schooling organization to produce specific outcomes is partially related to financial resources allocated by a society that is more concerned with lower tax rates than with mastery. Non-masters, i.e. students in need of additional assistance are determined therefore, not on the basis of any absolute standard but rather on the basis of resource availability. This leads to setting standards that permit a known percentage of students to receive special assistance.

This 'Relative Standard' approach is used by the State Education Department of New York in setting boundaries for remedial emphasis on PEP tests as well as its use of

resources based on CAR reports. The lowest portion of a population is served based on the assumption that the entire system will best be served by focusing its use of limited resources on those in greatest need. There has been no attempt to set an 'absolute value' as the target for either individual students or school systems on New York State mandated tests.. Such a target could only be arbitary and would be subject to continual controversey.

Given this 'relative cut-off score' argument, it becomes obvious that CRT application in grade level promotional decisions is currently the only truely defensible use of CRT's. Given the limitations of arbitrarily established absolute cut-offs, CRT's have limited value in setting graduation and licensure decisions.

Reliability: Unlike validity, reliability is a test characteristic. Reliability is an index of the instruments (test) ability to repeatedly produce the same result. Reliability for norm-referenced tests is based on how closely the same score for each individual can be replicated. The closer a test comes to repeating the same score for each examinee on two different administrations, the higher the reliability. In CRT's however, one is concerned only with the test's ability to replicate the master/non-mastery distinction. This is a qualitative

decision compared with the quantitative decision required in norm-referenced tests.

Reliability is calculated for norm-referenced tests by administering alternate forms of the test to the same population and then correlating scores of individuals on the two administrations. A second approach compares the score on one-half of the test with a score on the other half (split-half or internal consistency reliability). Because CRT reliability is concerned only with replicating the mastery/non-mastery distinction, a slightly different computational procedure is applied with either the alternate forms or split half approach. Although it is generally recognized that limited student response variability on CRT's will result in lower reliability estimates. Kane (1986) has demonstrated that reliabilities below .50 must be viewed with caution.

Discrimination is an item characteristic rather than a test characteristic. A discrimination value describes the frequency that masters respond correctly to a single test item and that non-masters respond incorrectly. Items that non-masters get correct as often as masters do fail to discriminate between the two populaltions. The discrimination index is a classical test item characteristic that is applicable to CRT items. Modification in interpretation is, however required. Normally, in

norm-referenced testing, one looks for item discrimination indices of .30 and above. That is, the ratio of high scoring students producing a correct response to the ratio of low scoring students getting that item correct is .30 or higher. Although still a good general rule of thumb to follow in CRT development, it is also important to examine each test item carefully to determine how the item relates to the mastery distinction. Given the criteria of item value, discrimination indices with lower values may be retained in CRT applications.

Degree of Difficulty is another test item characteristic. The degree of difficulty describes how often examinees are likely to respond correctly to the item. This is another of the classical test item characteristics that is applicable to CRT items. A degree of difficulty may be computed for each test item. Although traditionally a 50% error rate is considered desirable in norm-referenced test construction, CRT's tend to focus on a degree of difficulty that is more sensitive near the cut-off score. That is, individual test item difficulty should be determined in conjunction with the cut-off score. If the final cut-off range is in the 70% to 80% area, then item difficulty should be set near these values.

Field Testing

Each of the CRT statistics described above must be established with a sample of the population for whom the test is intended. The determination of a suffucient sample is often a dilemma. The more closely a sample approximates that of the total population, the more accurate are the estimates of population characteristics. There is, however a point of diminishing returns. There is a point at which increasing sample size contributes to increased accuracy of the statistical estimates only minimally. My own rule of thumb is the larger of 20% or 100 examinees with a representative corss section of the total population.

Individual test times are examined for both degree of difficulty and discrimination characteristics. Groups of items to be administered as a 'Test' are examined for reliability. Because reliability is a function of test length, short tests to be used with only one or two objectives are not subjected to reliability analyses.

It may be necessary to field test CRT's on repeated occasions if initial use produces undesirable statistical properties. The possibility of repeated testing should be built into any plan for development of a CRT. It is also desirable to use up to two or three times as many items as needed in field testing in recognition of the fact that some items will be discarded because of their statistical properties.

Administration

Test items will be packaged for administration much like existing standardized tests. In fact, existing standardized test forms may be used as models for the development of CRT's.

Summative CRT's are administered at the end of the instructional process in a secure testing situation similar to standardized test administrations. Frequency of summative testing for critical decision making, i.e. placement of individual students is an issue to be considered. Will major placement decisions occur at every grade level or only at specific grade levels where remedial resources are concentrated for more effective utilization? The frequency of summative testing and critical decision points will depend on the availability of remedial resources. Districts have used a variety of approaches to establishing Gates for uninterupted promotion, e.g. gates at grades K, 2, 5 and 7; every grade level, grade four only, etc..

When to begin summative testing is yet another issue. Developmentalists and proponents of Whole Language will argue that formal testing should not begin until grades 3 or 4. Measurement specialists and proponents of accountability will argue that formal testing should begin in Kindergarden.

This decision will depend on value judgments within the district and will likely vary from district to district.

Administration of formative CRT's will depend on several factors. If curricular objectives are closely aligned with a specific textbook or program sequence, then CRT's may be administered as unit tests. If curricular objectives are independent of program, i.e. teachers have the freedom to use any program they wish to achieve course objectives, then formative CRT's may be administered either as each objective is taught or at given intervals, eg. five weeks, quarterly, etc.

Management

Although formative CRT's are designed primarily for teacher use in instructional decision making within individual classrooms, some form of district-wide management system can increase the effectiveness of formative CRT's. Distribution, scoring and record keeping of formative CRT's can become a logistical nightmare unless a management system is created. A management system involves grouping of test items into some type of package so that teachers are not pulling out one set of 4 test items each time an objective is completed. Packaging test items implies some form of either Pacing (the rate at which instruction is to occur) or Sequencing (the order in which objectives are to be presented). Although

packaging test items and setting a testing schedule reduces teacher flexibility, administsrators will be wise to discuss the trade-offs with teachers.

The administration of formative CRT's will also depend on the nature of course objectives; that is, are objectives developmental or at higher, more generalizable levels. Developmentally stated objectives will require more testing than a few general level objectives.

CRT's at the primary level present a problem in that students at this level cannot be expected to transfer answers to machine scorable response sheets. Either hand scoring of tests must be done or special forms must be printed to accommodate student responses on the test pages. The number of test items that can be accommodated on each page for testing at the primary level is limited thereby creating a logistical problem for storage and distribution of test forms. NCS (Sheppard, 1986) has developed a generic machine scorable response sheet that facilitates immediate turn around for test results within the building through use of scanners, microcomputers and printers at the building level. These generic response sheets can be used at any level.

Results of CRT Application

In 1978, Hambleton (1978) stated that in theory, "... objectives based programs were designed to:

1. define curricula in terms of objectives.
2. use these objectives to drive instruction.
3. provide on-going evaluation data for instructional decisions." p. 280

He also concluced that hard evidence in support of this theory was in short supply.

Today, there appears to be no shortage of evidence to support the use of CRT technology in conjunction with objectives based programs. Reseach findings (Abrams, 1985; Barber, 1979; Conner, et. al. 1985; Conyers, et. al., 1985; Fuchs, et. al., 1986; Guskey and Gates, 1986; Hyman, 1979 and Mevarech, 1985) report that:

1. instruction directed at specifically defined behaviors (objectives/competencies) is far more effective than global instruction.
2. frequent, formal monitoring of students with CRT's aligned with curricular objectives/competencies is superior to teacher judgment of student progress.
3. student progress tied directly to specific objectives encourages more effective utilization of instructional resources.
4. parent involvement increases with the precise information made available in objectives based programs

accompanied by frequent CRT assessment and reporting of progress.

5. student knowledge of progress, accomplishments and expectations is increased with objectives based programs and CRT assessement.

These conclusions are derived from research summaries, meta-analysis of research reports and individual district summaries of student gains. There appears to be little doubt that the formative use of CRT's in well defined objectives based programs contributes to student achievement gains as measured by a variety of indices including the traditional 'Norm-Referenced' standardized test.

Not all reviews are so overwhelmingly positive. A dissenting review by Slavin, (1987) concludes that the claims of Mastery Learning proponents are highly exaggerated. His analysis of highly selective studies reveals at best very minor gains and he questions weather these gains are the result of more time on task for some students or the monitoring function. Obviously, frequent monitoring has led to more efficient remediation and the recent development of more alternatives in remedial efforts has increased their effectiveness. Weather increased student gains come from more time on task or frequent monitoring seems irrelevant. The fact is that frequent monitoring with formal tests has led to increased achievment of students.

References

Abrams, J. D. (1985). Making outcome-based education work, Educational Leadership, 43(1). 30-32.

Arlin, M. (1984). Time, equity, and mastery learning, Review of Educational Research. 54(1). 65-86.

Barber, C. (1979) Training principals and teachers for mastery learning. Educational Leadership. 37(3). 126-127.

Berk, R. A. (1980). A consumers' guide to criterion-referenced test reliability. Journal of Educational Measurement, 17(4). 323-349.

CIMS, (1986). Comprehensive Instructional Management System: Elementary Mathematics. NYS Syllabus, State Education Depatment of New York State: Albany, NY.

Conner, K., Hairston, J., Hill, I.; Kopple, H,, Marshell, J., Scholnick, K. & Schulman, M. (1985). Using formative testing at the classroom, school, and district level. Educational Leadership, 43(2). 63-67.

Conyers, J. G.; Andrews, K. & Marzano, R. J. (1985). Developing district made criterion referenced tests; a standard of excellence for effective schools. Education, 106(2). 141-149.

Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis. Exceptional Children, 53(3). 199-208.

Fuchs, L. S., Fuchs, D. & Tindal, G. (1986). Effects of mastery learning procedures on student achievement. Journal of Educational Research, 79(5). 286-291.

Geeslin, D. H., (1985). A survey of pupil opinion concerning learning for mastery. Education, 105(2). 147-150.

Glass, G. V. (1978). Standards and criteria. Journal of Educational Measurement, 15(4). 237-261.

Guskey, T. R. & Gates S. L. (1986). Synthesis of Research on the effects of mastery learning in elementary and secondary classrooms, Educational Leadership, 43(8). 73-80.

Haladyna, T. M. & Roid, G. H. (1983) A comparison of two approaches to criterion referenced test construction. Journal of Educational Measurement, 20(3). 271-282.

Hambleton, R. K. (1978). On the use of cut-off scors with criterion referenced tests in instructional settings. Journal of Educational Measurement, 15(4). 277-290.

Hambleton, R. K. & De Gruijter, D. N. M. (1983). Application of item response models to criterion-referenced test item selection. Journal of Educational Measurement, (20)4. 355-366.

Hambleton, R. K., Graig, N. M. & Simon, R. (1983). Determining the lengths for criterion referenced tests. Journal of Educational Measurement, 20(1), 27-38.

Haertel, E. (1985). Construct validity and criterion-referenced testing. Review of Educational Research, 55(1). 23-46.

Hambleton, R. K., Swaminathan, H., Algina, J. & Coulson, D. B. (1978) Criterion-referenced testing and measuremeent: a review of technical issues and developments. Review of Educational Research, 48(1). 1-47.

Holmes, C. T. & Matthews, K. M. (1984). The effects of nonpromotion on elementary and junior high school pupils: a meta-analysis. Review of Educational Research, 54(2). 225-236.

Hyman, J. S. & Cohen, S. A. (1979). Learning for mastery: ten conclusions after 15 years and 3,000 schools. Educational Leadership, 37(3). 104-108.

Kane, M. T. (1986). The role of reliability in criterion-referenced tests. The Journal of Educational Measurement, 23(3). 221-224.

Lathrop, R. L. (1986). Practical strategies for dealing with unreliability in competency assessment, Journal of Educational Research, 70(4). 234-237.

Levine, D. U. & Levine R. F. (1986) Accountability implications of effective teaching competencies, Education and Urban Society, 18(2). 230-241.

Linn, R. L. (1978). Demands, cautions and suggestions for standard setting. Journal of Educational Measurement, 15(4). 301-308.

Mevarech, Z. R., (1985). The effects of cooperative mastery learning strategies on mathematics achievment. Journal of Educational Research, 78(6). 372-377.

ORBIT (1984). Objectives-Referenced Bank of Items & Tests: Technical Bulletin 1, Monterey, CA: CTB/McGraw-Hill.

Peterson, S. E., Degracie, J. S. & Ayabe, C. R. (1987). A longitudinal study of the effects retention/promotion on academic achievement. American Educational Research Journal. 24(1). 107-118.

Popham, J. W. & Husek, T. R. (1969) Implications of criterion-referenced measurement. Journal of Educational Measurement, 6(1). 1-9.

Ruff, T. P. (1985). Avoiding hucksters, fakirs, gurus and other academic heretics. The Clearing House, 59(1). 16-18.

Secolsky, C. (1987). Attitudinal and interpertive influences on difficulty ratings in setting standards for high school minimum competency tests. Journal of Educational Research, 80(4). 227-232.

Shaycoft, M. F. (1979) Handbook of Criterion-Referenced Testing. Garland STPM Press: NY, NY.

Shepard, L. (1980). Standard setting issues and methods. Applied Psychological Measurement, 4(4). 447-467.

Sheppard, D., (1986). Laurel schools committed to automated instructional management. Educational Technology Report, National Computer Systems: Fall. 5-6.

Slavin, R. E. (1987). Mastery Learning reconsidered. Review of Educational Research, 57(2). 175-213.

Stiggins, R. J. (1985). Improving assessment where it means the most: in the classroom. Educational Leadership, 43(2). 69-73.

Tindal, G.; Fuchs, L. S.; Fuchs, D.; Shinn, M. R.; Deno, S. L. & German, G. (1985). Empirical validation of criterion-referenced tests. Journal of Educational Research, 78(4). 203-209.

Van Der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the angoff and nedelsky techniques of standard setting. Journal of Educational Measurement, 19(4). 295-308.

Witthuhn, J., (1986). Instructional management: a tool for increasing productivity. T.H.E. Journal, March. 94-96.